Frequency Domain Linear Prediction for Speech and Audio Applications

Sriram Ganapathy
IDIAP Research Institute
P.O. Box 592
CH-1920, Martigny
Switzerland
ganapathy@idiap.ch

Thesis Proposal
November 21, 2007

Submitted to:
Ecole Polytechnique Fédérale de Lausanne (EPFL)

Supervisor:
Prof. Hynek Hermansky
IDIAP Research Institute
Ecole Polytechnique Fédérale de Lausanne (EPFL)

1 Motivation

Wide-band audio coding algorithms have become very popular due to emerging digital audio applications for network, wireless and multimedia computing systems. Digital storage of audio signals, which results in reduced size and higher quality is commonplace in compact discs, digital video discs and has became a standard in most of audio/video portable electronic devices.

Current state-of-the-art coding systems are application-specific. Although, it is hard to objectively measure quality across available audio codecs, it can be safely claimed that there is no generic codec that performs sufficiently well for all potential applications under all channel conditions. The state of the art audio systems are developed as a combination of several coding techniques and codec modes to meet all requirements. Therefore, there is a strong requirement to encode all types of signal using the same analysis model.

In this proposal, our objective is to develop a novel wideband generic speech/audio coding technique that will take advantage of long temporal contexts. Unlike conventional state-of-the-art systems based on short-term processing of the signal, we will focus on the dual method of processing short frequency bands over long temporal segments of the signal.

Long term temporal context has been successfully employed for a wide range of areas in speech and audio processing like speech recognition [12, 20, 21], speech enhancement [15], speech coding [35] etc. The modulation spectrogram [18, 25], which forms an alternative for the convention spectrogram, has also been widely used as a speech analysis tool. Also, the duality of the spectral-temporal modelling has been explored in [13].

Here, we propose to exploit the predictability of slowly varying amplitude modulations for coding speech and audio signals. Our approach is based on representing Amplitude Modulating (AM) signal using Hilbert envelope estimate and Frequency Modulating (FM) signal using Hilbert Carrier. With a small approximation error the Hilbert envelope (squared magnitude of the analytic signal) can be shown to be the squared AM envelope [30]. Since we apply LP techniques to represent modulations in the time-frequency domain, we call this approach Frequency Domain Linear Prediction (FDLP) [28]. The proposed FDLP technique originates from the duality in modeling the temporal envelopes and power spectrum [26].

The remaining part of this proposal is organized as follows. We describe the state of the art audio coding schemes in Section 2. Scope of the proposed research is discussed in Section 3. The preparatory work done is detailed in Section 4 and the schedule for the future work is sketched in Section 5.

2 State of Art

During the last decade, CD-quality digital audio has essentially replaced analog audio. Emerging digital speech and audio applications for network, wireless, and multimedia computing systems face a series of constraints such as reduced channel bandwidth, limited storage capacity, and low cost. These new applications have created a demand for high-quality digital audio delivery at low bit rates. In response to this need, considerable research has been devoted to the development of algorithms for perceptually transparent coding of high-fidelity digital audio. As a result, many algorithms have been proposed, and several have now become international and/or commercial product standards.

2.1 Speech coding

In order to simplify the description of speech codecs they are often broadly divided into three classes - waveform codecs, source codecs and hybrid codecs [31]. Typically waveform codecs are used at high bit rates, and give very good quality speech. Source codecs operate at very low bit rates, but tend to produce speech which sounds synthetic. Hybrid codecs use techniques from both source and waveform coding, and give good quality speech at intermediate bit rates.

Waveform codecs attempt, without using any knowledge of signal production, to obtain a reconstructed signal whose waveform is as close as possible to the original. Generally, they are low complexity codecs which produce high quality speech at rates above about 16 kbits/s. When the data rate is lowered below this level, the reconstructed speech quality degrades rapidly. The simplest form of waveform coding is Pulse Code Modulation (PCM), which merely involves sampling and quantizing the input waveform.

Source coders operate using a model of how the source was generated, and attempt to extract, from the signal being coded, the parameters of the model. It is these model parameters which are transmitted to the decoder. Source coders for speech are called vocoders.

Hybrid codecs attempt to fill the gap between waveform and source codecs. Among the popular current speech codecs belong Code-Excited Linear Prediction (CELP) [11], which is an umbrella term for a family of techniques that quantize the residual obtained from linear prediction using Vector Quantization (VQ). The CELP technology has been adopted in Adaptive Multi-Rate (AMR) technique [23], known as one of the best current audio data compression scheme for speech coding. AMR was standardized by the 3rd Generation Partnership Project (3GPP) [3] and is now widely used in Global System for Mobile communications (GSM).

2.2 Audio coding

Just as the audio CD stimulated a tremendous boom for the record industry, the booming popularity of the Internet has opened up new ways for the promotion and distribution of music to consumers. Furthermore, using IP network as a new service platform, telecommunication providers foresee new opportunities, as services over IP reduce costs, maximize bandwidth efficiency and introduce new possibilities for the customer. As a consequence, one may think of other services such as live audio and video streaming applications (e.g., radio and TV broadcast over IP, multi-cast of a lecture, etc.).

Due to new potential applications and requirements, there arises a need for new wideband speech and audio coding technologies. The first international standard for digital compression of high-fidelity audio was delivered by the Moving Pictures Experts group (MPEG) [4], and later adopted by ISO/IEC. MPEG-1, the first generic audio compression standard (standardized by MPEG in 1992) is not based on modeling of the speech production system. It belongs to the class of sub-band (or transform) coders, where the compression is achieved by exploiting perceptual limitations of the human auditory system. MPEG-1 offers several compression modes. Particularly, the architecture contains three layers of different output quality, increasing complexity and delay. The most popular architecture is **MP3**, MPEG-1 layer III architecture [32, 31].

Demand for high quality coding of multichannel data on reduced bit-rates made MPEG to adopt new versions of the architectures. MPEG-2 provides extensions to lower sampling rates and support backward compatibility of multichannel stereo modes. Third-generation MPEG audio coder is an extension of the MPEG-2 standard and is referred to as MPEG-2 **Advanced Audio Coder (AAC)** [7, 14, 22]. AAC nowadays represents the high quality end of MPEG-4 standard. AAC is completely different codec offering large improvement over MPEG-1 in terms of efficiency and quality.

AAC architecture became a base technology for MPEG-4 standard, where attention was focused on low bit-rate coding below 64 kbps per audio channel. MPEG-4 can be seen as set of coding tools offering many concepts, such as universality, or scalability. MPEG-4 High Efficiency AAC version 2 (HE-AAC v2), also known as AAC+ (v2) [1], is one of the most popular AAC coding formats. HE-AAC v2 is the combination of three technologies: AAC, Spectral Band Replication (SBR) and Parametric Stereo (PS), and has been standardized by the European Telecommunications Standard Institute (ETSI). It was also standardized by 3GPP. Most commonly used file formants are MP4 and M4A.

Adaptive Multi-Rate (AMR) technique, which has been widely used in GSM, has been extended for wideband speech coding applications (AMR-WB) [29]. It is also based on CELP algorithm (Algebraic CELP) and it spans bandwidth $50 - 7000$ Hz. AMR-WB is referred as G.722.2, in ITU-T standardization organization. Later, AMR-WB has been extended into AMR-WB+ [2] for encoding any kind of input signal. AMR-WB+ uses transform coding (with a technique known as TCX - transform-coded excitation) as an addition to the algebraic CELP algorithm exploited in AMR-WB. Incorporated transform coding largely improves generic audio coding. The selection of the mode between transform coding and algebraic CELP is able to provide good speech and audio quality with moderate bit-rates. AMR-WB+ has been developed by 3GPP for messaging and streaming services used by GSM, and for 3rd generation cellular services.

2.3 Codec quality comparisons

Subjective quality assessments, performing psychoacoustic measurements generated by experimental means, is the only approach providing sufficiently reliable results to evaluate different speech and audio compression techniques. There exist internationally standardized test methods for subjective measuring the quality of audio coding techniques. One of the most common test methodology is MUSHRA [8] (MUlti-Stimulus test with Hidden Reference and Anchor) standardized in ITU-R. On each trial in a MUSHRA test, the subject is presented with an uncompressed audio sample - Open Reference (OR). A score 100 is given to the quality of OR on the MUSHRA quality scale. The goal of the subject is to evaluate the quality of the same audio sample processed by an audio codec involved in the test as well as the uncompressed audio sample - Hidden Reference (HR), and to one or more degraded anchor samples, typically low-pass filtered at 7 kHz and 3.5 kHz.

However, subjective listening tests are time-consuming, expensive and impractical for everyday use. It is beneficial, especially for development purposes, to substitute the subjective listening tests with objective, computer-based methods. Perceptual Evaluation of Speech Quality (PESQ) [6] and Perceptual Evaluation of Audio Quality (PEAQ) [34] are standardized methods (standardized by ITU-R as P.862 and BS.1387, respectively) for objective assessment of speech and audio coders. These methods are able to significantly speed-up the development process of new codec.

Compression efficiency versus quality is not the only parameter to evaluate audio codecs. Other important features are scalability, robustness against channel errors (packet loss) and computational complexity.

2.4 Shortcomings of current audio coding techniques and future trends

State-of-the-art standardized codecs, where dominant roles are played by AAC and AMR technologies, perform well in specific tasks, but do not offer sufficient qualities across various input signals and channel conditions. Therefore, these systems usually comprise of several codecs used for encoding different types of inputs. The utilization of several codec modes also indicate the lack of a single analysis tool capable of modelling speech and audio signal. Therefore, in the area of speech and audio processing, there is a growing trend towards a signal independent analysis tool. Some earlier attempts like sinusoidal modelling [33], energy operator [27] were made in this direction. But, none of them were significantly successful for all types of input conditions.

With huge expansion of digital technologies, it is not easy to predict acceptance of new codecs. Commercial success or failure of the codecs depends not only on the technology but also on issues such as hardware compatibility and upgrade-ability, costs of new technology, licensing terms, and other. However, the future of audio codecs looks bright with the expansion of digital technologies.

3 Scope of Proposed Research

From the point of view of current and future services, there is a need for high-quality wideband (16 - 48 kHz) generic audio coding on low and medium data rates (8 - 30 kbps). In this proposal, our objective is to develop new generation of generic audio coding technique, based on the principles of Frequency Domain Linear Prediction. The main idea exploited in the proposed codec will be the use of information from relatively long temporal context (few hundreds of ms). The goal is to develop the codec which will be capable of encoding any kind of input speech and audio signal. The codec must also be scalable across a wide range of bit-rates without any additional computational burden. Another important parameter of the proposed codec (besides the compression quality) will be its robustness against channel errors. Variable Bit-Rate (VBR) encoding based on the input content and channel conditions and multi channel coding form other important objectives for the proposed codec.

In addition to the proposed speech and audio coding work, we would also explore the applicability of FDLP as features for ASR. Although the technique has been employed for small vocabulary tasks [12], we will extend the work for larger databases as well as noisy input conditions. In particular, we will utilize FDLP as a time-frequency analysis tool similar to the modulation spectrogram [18]. In this manner, the research would prove useful to a broader scientific community in speech and audio engineering.

4 Preparatory Work : FDLP Based Codec

The novelty of the proposed audio coding approach is the employment of FDLP method to parameterize Hilbert envelope (squared magnitude of an analytic signal) of the input signal [22, 12]. FDLP can be seen as a method dual to Temporal Domain Linear Prediction (TDLP). In the case of TDLP, the AR model approximates the power spectrum of the input signal. FDLP fits an AR model to the squared Hilbert envelope of the input signal. Using FDLP, we can adaptively capture fine temporal nuances with high temporal resolution while at the same time summarize the signal's gross temporal evolution in time scales of hundreds of milliseconds. In our system, we employ the FDLP technique to approximate the temporal envelope of sub-band signal in QMF sub-bands.

4.1 Structure of the codec

The full-band input signal is decomposed into sub-bands using QMF analysis. In each sub-band, FDLP is applied and LSFs (Line Spectral Frequencies) representing the sub-band Hilbert envelopes are quantized. The residuals (sub-band carriers) are processed using DFT and corresponding spectral parameters are quantized. In the decoder, spectral components of the sub-band carriers are reconstructed and transformed into time-domain using inverse DFT. The reconstructed FDLP envelopes (from LSF parameters) are used to modulate the corresponding sub-band carriers. Finally, the inverse QMF block is applied to reconstruct the full-band signal from frequency sub-bands.

In following subsections, we describe each of these blocks present in the encoder in more detail. The blocks at the decoder perform the inverse operation of those at the encoder.

4.2 Encoder

Graphical scheme of the encoder is given in Figure 1. The important blocks present are:

4.2.1 QMF Analysis

The input audio signal is decomposed into 32 non-uniform critically sampled sub-bands. For this purpose, the input audio signal is first decomposed into 64 uniform sub-bands using a uniform QMF decomposition. The 64 uniform QMF bands are then merged to obtain 32 non-uniform bands to approximate the bark scale.

4.2.2 FDLP processing

The technique of FDLP, described in [12], is applied on each critically sampled sub-band. Specifically, the signal is predicted in the DCT domain and the

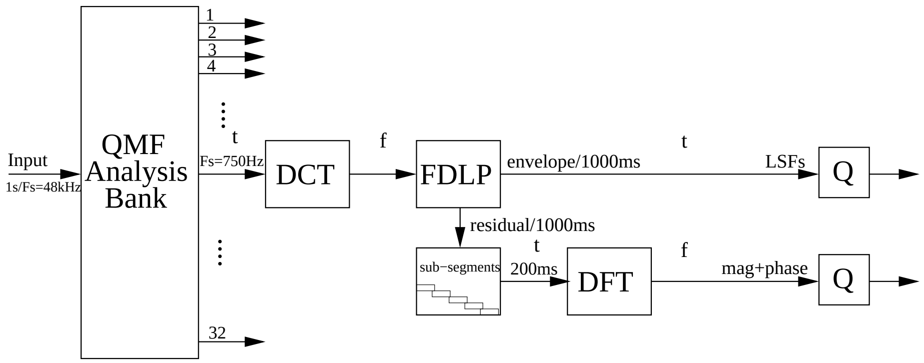

Figure 1: Schematic diagram of the encoder (f - frequency domain, t - time domain, QMF - Quadrature Mirror Filter bank, DCT - Discrete Cosine Transform, FDLP - Frequency Domain Linear Prediction, DFT - Discrete Fourier Transform, Q - Quantization).

prediction coefficients form the FDLP envelope. The residual of the prediction forms the FDLP carrier signal. So the function performed by the FDLP block is to split the signal into two parts, namely an approximation part represented by the envelope coefficients and an error in approximation represented by the FDLP carrier.

4.2.3 Quantization of parameters

Quantization of LSFs: The LSFs corresponding to the AR model in a given frequency sub-band over the 1000 ms input signal are vector quantized.
Quantization of the magnitude components of the DFT transformed sub-segment residual: The magnitude spectral components are vector quantized. Since a full-search VQ in this high dimensional space would be computationally infeasible, the split VQ approach is employed. Although the split VQ approach is suboptimal, it reduces computational complexity and memory requirements to manageable limits without severely affecting the VQ performance.
Quantization of the phase components of the DFT transformed sub-segment residual: It is found that the phase components are uncorrelated across time. The phase components have a distribution close to uniform, and therefore, have a high entropy. Hence, we apply a uniform scalar quantization for the phase components. To prevent excessive consumption of bits to represent phase coefficients, those corresponding to relatively low magnitude spectral components are not transmitted, i.e., the codebook vector selected from the codebook is processed by adaptive thresholding in the encoder as well as in the decoder. Only the spectral phase components whose magnitudes are above the threshold are transmitted. The threshold is adapted dynamically to meet a required number of spectral phase components (bit-rate).

4.2.4 Physco-acoustic Modelling

Although the basic technique achieves good quality of the reconstructed signal, there is a need for improving the coding efficiency. The auditory masking properties of the human ear provide an efficient solution for quantization of a

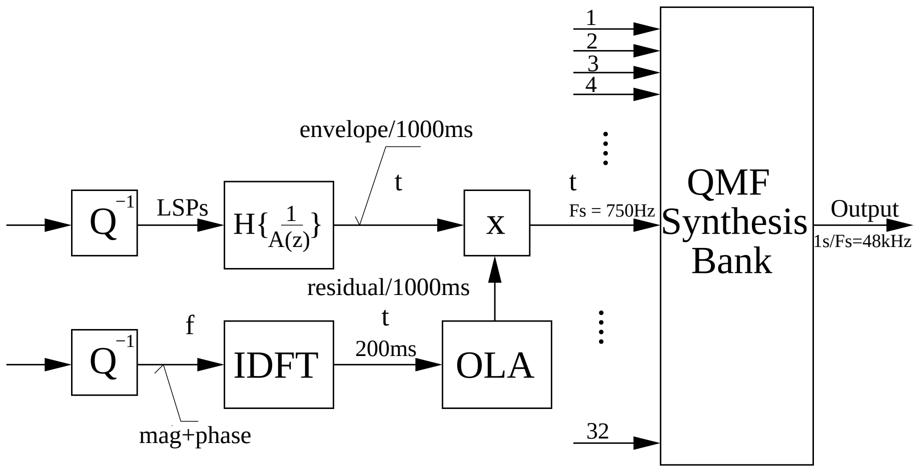

Figure 2: QMF-FDLP decoder structure (f - frequency domain, t - time domain).

signal in such a way that the audible distortion is minimized. In particular, temporal masking is a property of the human ear, where the sounds appearing within a temporal interval of about 200 ms after a signal component get masked [24]. In our experiments, a first order forward masking model of the human ear is implemented and informal listening experiments using additive white noise are performed to obtain the exact noise masking thresholds. Subsequently, this masking model is employed in encoding the sub-band FDLP carrier signal. Application of the temporal masking in the FDLP codec results in a bit-rate reduction of about 10% without degrading the quality.

4.2.5 Decoding

The decoder is shown in figure 2. In order to reconstruct the input signal, the carrier in each sub-band needs to be reproduced and then modulated by temporal envelope given by FDLP model.

The transmitted VQ codebook indices are used to select appropriate codebook vectors for the magnitude spectral components. Then, the adaptive threshold is applied on the magnitudes and the transmitted scalar quantized phase spectral components are assigned to the magnitudes lying above the adaptive threshold. The sub-band carrier is created in the time domain from its spectral magnitude and phase information. The Overlap-add (OLA) technique is applied to obtain 1000 ms residual signal, which is then modulated by the FDLP envelope to obtain the reconstructed sub-band signal. Finally, a QMF synthesis bank is applied on the reconstructed sub-band signals to produce the output signal.

4.3 Results

The qualitative performance of the FDLP codec is evaluated using Perceptual Evaluation of Audio Quality (PEAQ) distortion measure [34]. In general, the perceptual degradation of the test signal with respect to the reference signal is

bit-rate [kbps]	66	64	64
system	*FDLP*	*LAME*	*aacPlusv1*
ODG Scores	-1.11	-1.61	-0.77

Table 1: *Mean objective quality test results provided by PEAQ [34] for* 27 *files with mixed signal content from MPEG database for explorations in Speech and Audio Coding [10]*

measured, based on the ITU-R BS.1387 (PEAQ) standard. The output combines a number of model output variables (MOV's) into a single measure, the Objective Difference Grade (ODG) score. ODG is an impairment scale which indicates the measured basic audio quality of the signal under test on a continuous scale from -4 (very annoying impairment) to 0 (imperceptible impairment). The test was performed with 27 mixed signal content files from MPEG database for explorations in speech and audio coding [10] The results of objective quality evaluations are shown in Table 1.

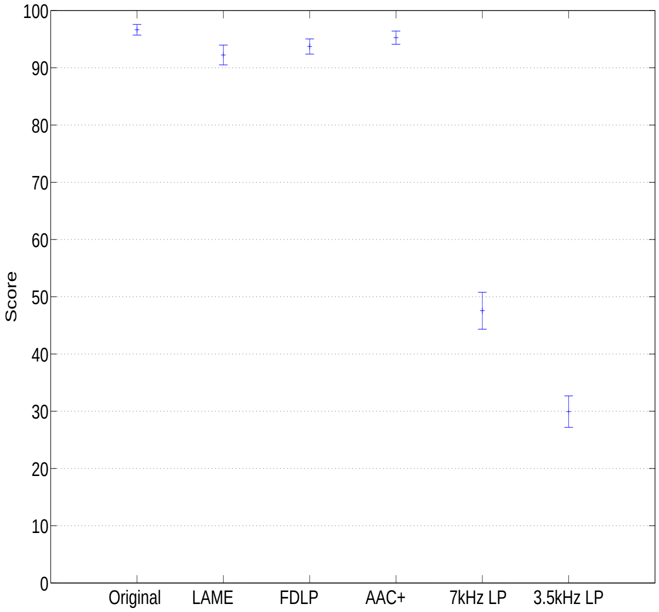

Figure 3: MUSHRA results for 8 mixed content files and 22 listeners for hidden reference (original), codecs (LAME, FDLP, aacPlus v1) and two anchors (7 kHz and 3.5 kHz low-pass filtered).

We evaluated the FDLP codec using the MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) method for subjective quality evaluation [8] with 8 audio samples from MPEG database for explorations in speech and audio coding [10] and 22 listeners. We compare the subjective quality of FDLP at 66 kbps, LAME - MP3 (MPEG 1, layer 3) at 64 kbps [5] and aacPlus v1 at ∼64 kbps [9]. The cumulative MUSHRA scores (mean values with 95% confidence) are shown in Figure 3. Based on subjective evaluation, we conclude that the FDLP codec performs better than LAME-MP3 and closely achieves subjective results of aacPlus v1 standard.

5 Schedule

The following are the summary of the steps in proposed research that are anticipated in the course of this work :

Task A Further extension of the current version of the codec:
The current version of the codec, built on processing relatively long temporal context, uses few conventional algorithms for data-rate compression. Currently, the codec achieves similar performances as the best state-of-the-art systems on medium data-rates even without exploiting entropy coding or psychoacoustic modeling in the frequency domain. These algorithms have shown to be the key techniques in state-of-the-art audio coding systems to achieve significant compression qualities.

A.1 Psychoacoustic modeling:
We will focus on traditional approaches based on psychoacoustic models to significantly reduce data-rates without compromising reproduction quality. State-of-the-art audio compression techniques deeply rely upon psychoacoustic models to optimize coding efficiency. Although, there exist many techniques in the field, they are developed for traditional systems based on processing short-term input segments. Therefore, such techniques need to be modified to be exploited in our codec. Once a frequency masking model is employed, we will deal with combining temporal and frequency masking models resulting in 2-D psychoacoustic model.

A.2 Entropy Coding:
Entropy coder is capable of compressing and recovering any kind of digital data without any loss. In general, it is used to match the actual entropy of the quantized transmitted values with the entropy of their representation in the channel bit-stream. Dynamic Huffman coding [36] can be one of the promising techniques to be exploited in the codec.

A.4 Spectral Echo :
Since the FDLP analysis technique cannot model spectral impulses (tonal signals), the dual of the temporal pre-echo appears in the reconstructed signal. We will develop techniques to shape the quantization noise in such a way that it is inaudible in the reconstructed signal.

A.4 Lower Latency :
The current FDLP based codec operates with a delay of 1000ms. Since some applications demand a lower latency, we would try to reduce the windowing to the order of 200ms without affecting the compression efficiency.

A.5 Block of quantization:
Quantization process in the codec is performed by simple scalar and

vector quantization. We will put an additional effort in this domain to utilize more efficient techniques like adaptive quantizations and transform domain coding techniques.

A.6 Variable bit-rate coding:
The goal is to develop a high scalable codec operating from low to medium bit-rates depending on input and channel conditions, i.e., utilizing bit-reservoir controlled by properties of input signal and channel conditions. This task is closely tied with the previous tasks (especially with psychoacoustic modeling).

A.7 Multi-channel audio coding:
As most of the state of art codecs have a multi-channel encoding option, we will also provide a multi-channel encoding facility in the FDLP based codec.

Task B Experimental comparisons:
In order to evaluate proposed audio compression technique, experimental comparisons with current state-of-the-art systems will be made. First, objective quality comparisons (PEAQ or PESQ) will be used for developments purposes. Subjective quality evaluations based on standardized approaches will be performed on challenging well-known benchmark databases.

Task C FDLP based features for ASR: Besides speech and audio coding, we would also focus on exploiting FDLP as features for ASR. Traditionally, short segments of the signal ($10 - 30$ ms) are used to derive short-term features for pattern classification in ASR [19]. However, a single frame of a short-term spectrum does not contain all the information that is necessary for decoding the lexical content of a given segment of speech. This is because the neighboring speech short term frames influence the short-term spectrum. Long temporal contexts have also been used for small vocabulary ASR tasks [12] and speech enhancement [16]. The utilization of FDLP features for larger databases will be explored in detail and resulting ASR performances will be compared with state of the art features. Developing noise robust features from FDLP will also be an area of focus, since long term processing has provided more robustness of ASR systems to certain kinds of environmental noise (for example reverberation noise [17]).

Task D Thesis writing.

5.1 Time Table

The following time table gives the approximate duration for the tasks detailed previously. Even though the table indicates the tasks as sequential, some overlap in time is obviously anticipated.

Task A: 15 months

Task B: 3 months

Task C: 8 months

Task D: 4 months

Total : 30 months.

6 Publications and Research Reports

- P. Motlicek, H. Hermansky, S. Ganapathy, H. Garudadri. "Non-Uniform Speech/Audio Coding Exploiting Predictability of Temporal Evolution of Spectral Envelopes", Proceedings of TSD 2007, LNCS/LNAI series, Springer-Verlag, Berlin, pp. 350-357, September 2007.

- P. Motlicek, S. Ganapathy, H. Hermansky, and Harinath Garudadri, "Frequency Domain Linear Prediction for QMF Sub-bands and Applications to Audio coding", Proceedings of MLMI 2007, LNCS Series, Springer-Verlag, Berlin, 2007.

- S. Ganapathy, P. Motlicek, H. Hermansky, H. Garudadri, "Temporal Masking for Bit-rate Reduction in Audio Codec Based on Frequency Domain Linear Prediction", *Tech. Rep., IDIAP*, RR 07-48, 2007.

- P. Motlicek, S. Ganapathy, H. Hermansky, and Harinath Garudadri, "Non-uniform QMF Decomposition for Wide-band Audio Coding based on Frequency Domain Linear Prediction ", *Tech. Rep., IDIAP*, RR 07-43, 2007.

- P. Motlicek, S. Ganapathy, H. Hermansky, and Harinath Garudadri, "Scalable Wide-band Audio Codec based on Frequency Domain Linear Prediction", *Tech. Rep., IDIAP*, RR 07-16, 2007.

References

[1] Enhanced aacplus General Audio Codec. *3GPP TS 26.401.*

[2] Extended amr wideband codec. *<http://www.3gpp.org/ftp/Specs/html-info/26290.htm>.*

[3] <http://www.3gpp.org/>.

[4] <http://www.mpeg.org/>.

[5] Lame mp3 codec. *<http://lame.sourceforge.net>.*

[6] Perceptual Evaluation of Speech Quality (pesq), an Objective Method for End-to-End Speech Quality Assessment of Narrowband Telephone Networks. *ITU-T Rec. P.862.*

[7] Generic Coding of Moving Pictures and Associated Audio: Advanced Audio Coding. *ISO/IEC JTC1/SC29/WG11 MPEG, International Standard ISO/IEC 13818-7*, 1997.

[8] Method for the Subjective Assessment of Intermediate Audio Quality. *ITU-R Recommendation BS.1534*, 2001.

[9] Coding of audio-visual objects part 3: Audio. *ISO/IEC Int. Std. 14496-3*, 2003.

[10] Framework for Exploration of Speech and Audio coding. *ISO/IEC JTC1/SC29/WG11*, 2007.

[11] B.S. Atal and M.R. Schroeder. Stochastic Coding of Speech at Very Low Bit Rates. *Proceedings of Int. Conf. on Comm.*, pages 1610–1613, 1984.

[12] M. Athineos and D. Ellis. Frequency Domain Linear Prediction for Temporal Features. *Automatic Speech Recognition and Understanding Workshop IEEE ASRU*, pages 261–266, 2003.

[13] Marios Athineos, Hynek Hermansky, and Daniel P.W. Ellis. PLP^2: Autoregressive Modeling of Auditory-like 2-D Spectro-Temporal Patterns. *IDIAP RR 04-60*, 2004.

[14] K. Brandenburg, O. Kunz, and A. Sugiyama. MPEG-4 Natural Audio Coding. *Signal Processing: Image Communication*, 15(4):423–444, 2000.

[15] T.H. Falk, S. Stadler, W.B. Kleijn, and Wai-Yip Chan. Noise suppression Based on Extending a Speech-Dominated Modulation Band. *Proceedings of Interspeech*, pages 970–973, 2007.

[16] D. Gelbart and N. Morgan. Evaluating long-term spectral subtraction for reverberant asr. *IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU*, pages 103 – 106, 2001.

[17] D. Gelbart and N. Morgan. Double the trouble: Handling noise and reverberation in far-field automatic speech recognition. *Proc. Int. Conf. Spoken Language Processing*, pages 2185–2188, 2002.

[18] Steven Greenberg and Brian E.D. Kingsbury. The modulation spectrogram: In pursuit of an invariant representation of speech. *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, pages 1647–1650, 1997.

[19] H. Hermansky. Perceptual linear prediction (plp) analysis of speech. *Journal of Acoust. Soc. of Amer.*, 87(4):1738–1752, 1990.

[20] H. Hermansky and P. Fousek. Multi Resolution RASTA Filtering for tandem Based ASR. *Proceedings of Interspeech*, pages 361–364, 2005.

[21] H. Hermansky and S. Sharma. TRAPS - Classifiers of Temporal Patterns. *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, pages 1543–1546, 1998.

[22] J. Herre and B. Grill. Overview of MPEG-4 Audio and its Applications in Mobile Communications. *International Conference on Signal Processing*, pages 11–20, 2000.

[23] K. Jarvinen. Standardization of the Adaptive Multi-Rate Codec. *Proceedings of European Signal Processing Conference*, pages 167–170, 2000.

[24] Walt Jesteadt, Sid P. Bacon, and James R. Lehman. Forward masking as a function of frequency, masker level, and signal delay. *Journal of Acoust. Soc. of Amer.*, 71(4):950–962, 1982.

[25] B.E.D. Kingsbury, N. Morgan, and S. Greenberg. Robust Speech Recognition using the Modulation Spectrogram. *Speech Communications*, 25:117–132, 1998.

[26] R. Kumerasan and A. Rao. Model-based Approach to Envelope and Positive Instantaneous Frequency Estimation of Signals with Speech Applications. *Journal of Acoustical Society of America*, 105:1912–1924, 1999.

[27] P. Maragos, J.F. Kaiser, and T.F. Quatieri. On amplitude and frequency demodulation using energy operators. *IEEE Transactions on Signal Processing*, 41(4):1532 – 1550, 1993.

[28] P. Motlicek, H. Hermansky, H. Garudadri, and N. Srinivasamurthy. Speech Coding Based on Spectral Dynamics. *Proc. of Text Speech and Dialogue 2006, LNCS/LNAI series, Springer-Verlag*, pages 471–478, 2006.

[29] Jari Mkin, Bruno Bessette, Stefan Bruhn, Pasi Ojala, Redwan Salami, and Anisse Taleb. AMR-WB+: A New Audio Coding Standard for 3rd Generation Mobile Audio Services. *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, pages 1109–1112, 2005.

[30] A.H. Nuttal and E. Bedrosian. On the Quadrature Approximation to the Hilbert Transform of Modulated Signals. *Proc. IEEE*, 54:1458–1459, 1966.

[31] T. Painter and A Spanias. Perceptual Coding of Digital Audio. *Proceedings of the IEEE*, 88:451–515, 2000.

[32] D. Pan. A Tutorial on MPEG Audio Compression. *IEEE Multimedia Journal*, 02(2):60–74, 2000.

[33] Xavier Serra. A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition. *Ph.D. Dissertation. Stanford University*, 1989.

[34] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, C. Colomes, M. Keyhl, G. Stoll, K. Brandenburg, and B. Feiten. Peaq - The ITU Standard for Objective. *ITU-R Rec. BS.1387*.

[35] M. S. Vinton and L. E. Atlas. Scalable and Progressive Audio Codec. *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, pages 3277–3280, 2001.

[36] J. S. Vitter. Design and Analysis of Dynamic Huffman Codes. *Journal of the ACM*, 34(4):825–845, 1987.